Enhancing Accuracy in Urban Air Quality Prediction: A Comparative Study of Predictive Algorithms for Air Pollutant Concentrations

Apoorva Verma[1]*, Dr. Leena Bhatia[2]

Abstract: Air pollution is a critical environmental issue affecting the health and well-being of populations globally. Thus, the accurate prediction of air pollution levels is essential for effective environmental management and public health interventions. Currently, research has focused on explaining the causes and temporal relationships among various factors affecting air quality, but very few studies have focused on the performance of wholesome forecasting models. This study conducted a comparative analysis of various algorithms employed to predict the concentration of diverse air pollutants in highly polluted cities in Rajasthan, India. Through extensive data collection and processing, this study evaluates the efficiency of eight different algorithms, namely, ARIMA, LR, SVM, Exponential Smoothing, Decision Tree Regressor, XG Boost, Random Forest and LSTM, for forecasting pollutant levels for the next five years, with the aim of identifying the most reliable and accurate models for air quality prediction in this specific geographical context. The findings suggest that the Decision Tree and XG Boost Provided the results with the highest accuracy of 42.39% and 42.38%, respectively. The findings also provide valuable insights for enhancing environmental monitoring and management strategies in regions facing severe air pollution challenges.

Keywords: Air Pollution, Air Quality, Algorithms, Comparative Analysis, India, Pollutants, Prediction models, Rajasthan

Introduction:

The term "pollution" describes tampering of the atmosphere with compounds that interfere with nature and, as a result, may severely harm human health. Air pollution is routinely associated with severe levels of air pollution in many regions of the world. The general well-being, sustainability of the environment, and public health are all seriously threatened by air pollution. Particulate Matter, Suspended Particulate Matter, Nitrogen Dioxide, Sulphur Dioxide, Ozone, Carbon Dioxide, and many more are the main air pollutants that are prevalent in the atmosphere [*23*].

Urban regions face serious environmental problems with air pollution, and The Land of Kings', Rajasthan, India, is no exception. Geographically, Rajasthan covers 10.4% of India's land, making it the largest state in the country. In Rajasthan, air pollution was a contributing factor in almost 1.13 lakh deaths in 2019.

[1]*Research Scholar, Rajasthan Technical University Kota, Assistant Prof., Dept. of MCA, Xavier Institute of Management and Informatics, St. Xavier's College, Jaipur, Orcid Id: 0000-0001-5741-2872. Email ID: v.apoorva1995@gmail.com*

[2]*Associate Professor, Department of Computer Applications, S.S. Jain Subodh PG College, Jaipur; Email ID: leenabhatia@hotmail.com*
Corresponding author: Apoorva Verma

Research Scholar, Rajasthan Technical University Kota, Assistant Prof., Dept. of MCA, Xavier Institute of Management and Informatics, St. Xavier's College,
Jaipur,   Orcid   Id: 0000-0001-5741-2872,   Email   ID:
v.apoorva1995@gmail.com

The reason for Rajasthan's high pollution levels, are fast urbanisation, yearly population growth (8.12 crore as per Census 2021), and its automobile fleet. According to the World Air Quality Report 2021, the major polluted cities in Rajasthan are Bhiwadi, Jaipur, Jodhpur, Udaipur, Pali, Kota, and Alwar. Thus, understanding and controlling air pollution in Rajasthan is one of the fundamental challenges, and to control the problem, the most important aspect is to understand upcoming trends in air quality, for which we need efficient models that can forecast air pollution with the utmost accuracy so that effective policies can be framed in advance to prevent the situation from worsening.

This paper aims to evaluate different time series and machine learning algorithms like Auto Regressive Integrated Moving Average (ARIMA), Linear Regression (LR), Support Vector Machine (SVM), Exponential Smoothing, Decision Tree Regressor, XG Boost, Random Forest and Long Short-Term Memory (LSTM) for forecasting the concentration of different prevalent pollutants for next 5 years. The evaluation was performed based on the calculated Error Rates. The remainder of this paper is organized as follows. The literature review conducted in this study is explained in Section 2. Section 3 describes the study's research techniques. The study's conclusion is provided in Section 5, followed by a discussion of the findings presented in Section 4.

Background Study:

As per latest reports Pollutants, including particulate

matter (PM), carbon monoxide (CO), nitrogen oxides (NOx), ground-level ozone (O3), and sulphur oxides (SOx), routinely surpass India's National Ambient Air Quality Standards [16]. Thus, it is necessary to address this problem. In recent years, many researchers have worked towards solving air pollution prediction problems. Artificial intelligence (AI), statistical, and physical forecasting models are the three categories of forecasting models. The study of meteorological conditions and weather events is the main focus of physical forecasting models. The goal of chemical transport models (CTM) is to build a mathematical model that explains meteorological and chemical processes in the atmosphere, with particular emphasis on the concentrations of air pollutants and their emission, transit, and mixing [12]. One of the main limitations of physical models is the large number of computations required, and the quality of these deterministic techniques depends on a large amount of data and information from sources of pollution. The physical model is simpler to compute and easier to utilize than the statistical model [17].

Accurate forecasting is difficult to obtain, and statistical models describe the relationships between variables based on statistical averages and feasibility. Previously, the autoregressive integrated moving average (ARIMA) model was used to effectively estimate various pollutants in Hyderabad [7]. Some other forecasting which was done using statistical model includes forecasting of levels of PM2.5 in Lahore [1], SO_2 and NO_2 in Surat [15], PM10 and SPM in Jodhpur [19] and many more. Statistical models are not only specific to forecasting, but they can also be used for finding correlations between various pollutants and other variables, such as meteorological factors of that study site, similar to the one done by Verma and Bhatia in Jaipur city. The results suggested that meteorological factors such as temperature, pressure, precipitation, and relative humidity significantly affect PM2.5 concentrations in Jaipur [22].

Several comparative studies conducted by different researchers have concluded that AI tools always perform better than physical forecasting and statistical methods [3]. For example, a comparative study done by Talamanova, I. *et.al.* performed a comparative study between LSTM, Exponential Smoothing and ARIMA, and the results suggested that Exponential Smoothing performed better [21]. Artificial intelligence (AI) solves complicated problems by imitating human vision, learning, and reasoning. One subset of artificial intelligence is machine learning (ML). Because of its high accuracy and robustness, machine learning is considered one of the finest techniques for pollution forecasting. It performs well in classification and regression series [24]. Machine learning methods have been used in various parts of the world to predict air pollution or the concentration of a particular pollutant in that region. Castelli, M. *et.al.* used a Support Vector Regressor (SVR) to predict air quality in California with

94.1 % accuracy [2], including the Gradient Boosting Method [20], bidirectional LSTM [13], Gaussian Naïve Bayes along with SVM [10], and Neural Network along with bidirectional LSTM [4].

Methodology:

The research methodology used in this study can be categorized as empirical research, specifically within the domains of Computer Science and Environmental Science. The methodology is divided into basic steps, in which the first step is data collection and pre- processing. The second step involved identifying the major pollutants prevalent in the highly polluted cities of Rajasthan. The third Step involves identifying trends and patterns in the collected and pre-processed air data. In the fourth step, we applied various statistical and machine-learning algorithms to forecast the concentration of air pollutants. Finally, we perform a comparative analysis of the applied algorithms. The methodology adopted for this study is explained in the chart below.
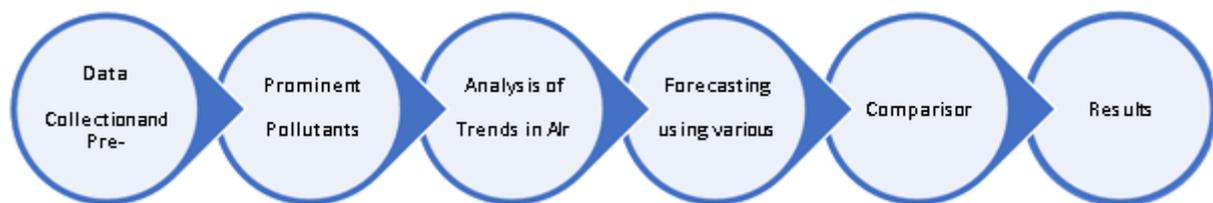

Fig 1: Flow Chart of the Study

1. Data collection and pre-processing

The data was gathered from the Central Pollution Control Board's (CPCB) official website, Central Control Room for Air Quality Management, which makes detailed real-time data available to the public [9].

Data of four major polluted cities, namely Bhiwadi,

Jaipur, Kota, and Udaipur, from January 2018 to July 2023. An image of the map displaying the study area is shown in Figure 2.
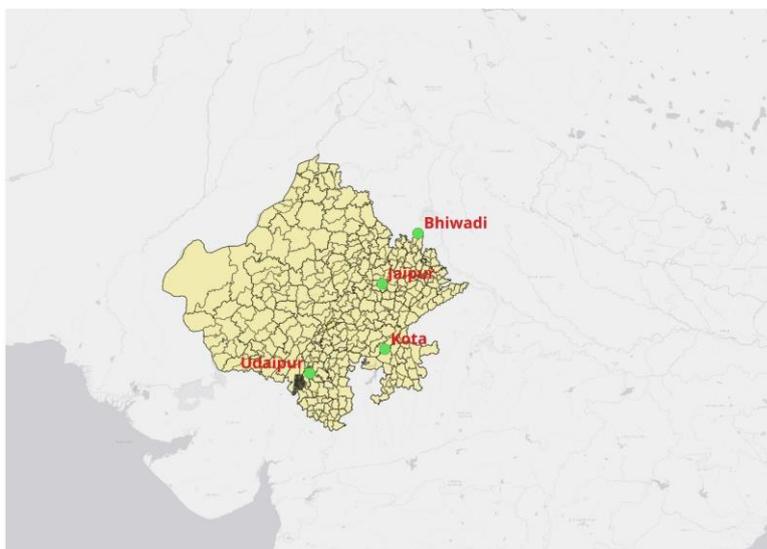
Fig 2: Map of Rajasthan showing area of study

The dataset contains all the component values present in the air, such as PM2.5, PM10, NO, NO2, NO$_X$, NH3, SO2, CO and Ozone. The data obtained are time-series data. After data collection, the data were pre-processed using the Python Panadas library. Data pre-processing in machine learning is the act of arranging and cleansing raw data so that it may be utilised to build and train models [14]. The refined data contained the following columns: Data, To Data, PM2.5, PM10, NO, NO2, NO$_X$, NH3, SO2, CO, Ozone, Area and City. The table below provides a description of the data.

Table 1: Dataset Description

Duration	Cities	Parameters	No. of Instances
1 January 2018 to 31 July 2023	4 (Bhiwadi, Jaipur, Kota, Udaipur)	9 (PM2.5, PM10, NO, NO2, NO$_X$, NH3, SO2, CO,	162252

2. Identification of prominent pollutants

After pre-processing the data, the data were grouped city-wise, and the most prevalent pollutants were identified. The graph below shows the most prevalent pollutants in the selected cities:
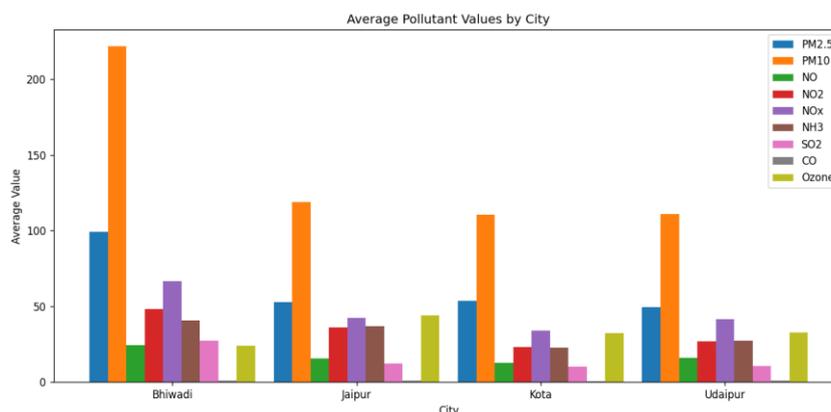

Fig 3: Average Pollutant Values City wise

From the graph, it is clear that the major pollutants prevalent in the four major polluted cities were PM2.5, PM10, NO$_X$ and Ozone. It is also seen the most prevalent pollutant was PM10.

3. Analysis of air quality trends

The continuous, multivariate time series of air quality data is made up of readings that represent a fixed measurement of time, with each reading dependent on the previous reading in some way [5]. Analysis of trends in air quality data provides valuable insights into long-term changes in air pollution levels [11]. Analysing trends helps in understanding whether the quality of air is improving or deteriorating over time in a particular geographical area. For our study, we used the Pandas and Matplotlib libraries of Python to calculate trends and visualize them, respectively. The trend analysis of the air quality in the four cities is shown in figure 3. It is clear from the graph that the air quality in all four chosen locations increased and decreased simultaneously.
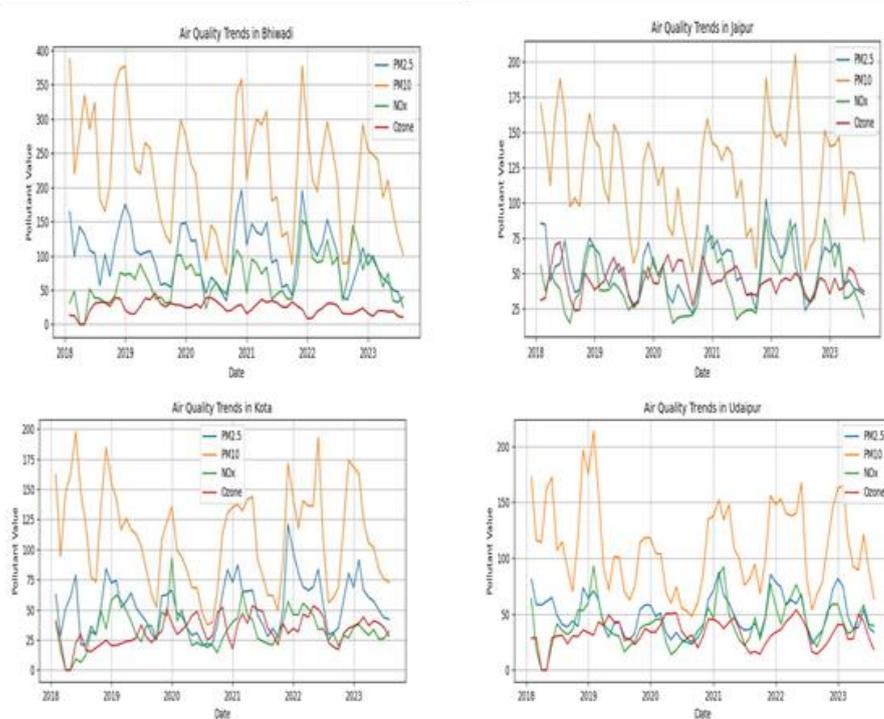
Fig 4: Air Quality Trends in Major Polluted Cities of Rajasthan

4. Forecasting using Various Algorithms:

Four major air pollutants, Particulate Matter 2.5 and 10, Oxides of Nitrogen and Ozone have been designated as needing constant monitoring at the aforementioned sites in Rajasthan, as of July 2023. After trend analysis, the next step was to forecast the concentrations of various pollutants in the selected cities using the selected algorithms. To forecast the concentration of the most prevalent air pollutants in the selected cities, we utilized both Time Series and Machine Learning Algorithms. The proposed architecture for forecasting is shown in figure below:
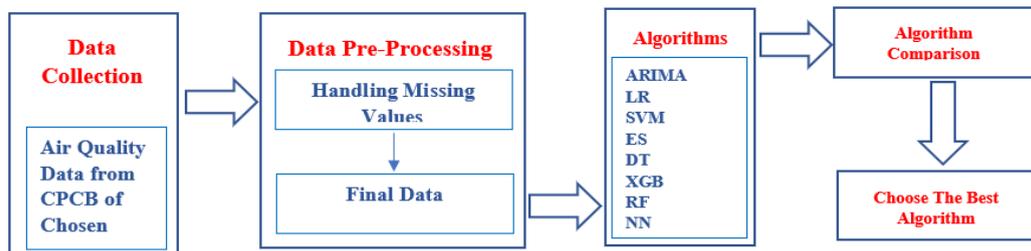
Fig 5: Methodology Used for Forecasting

Algorithms Used:

A. ARIMA:

The BOX-JENKINS modelling method was developed in the 1960s by the American academic BOX and British statistician JENKINS [1]. This is a comprehensive set of techniques for time-series analysis, prediction, and control. Depending on how challenging it is to extract seasonal effects, the ARIMA model is divided into two categories: a product seasonal model and a basic seasonal model. When a sequence exhibits both seasonal effects and short-term correlations and there is a complex relationship between the two, the sequence model can be fitted. For this study, the ARIMA model was developed

with p = d =q=1, and forecasting was performed for the next five years. The forecasting results are shown in the figure below:
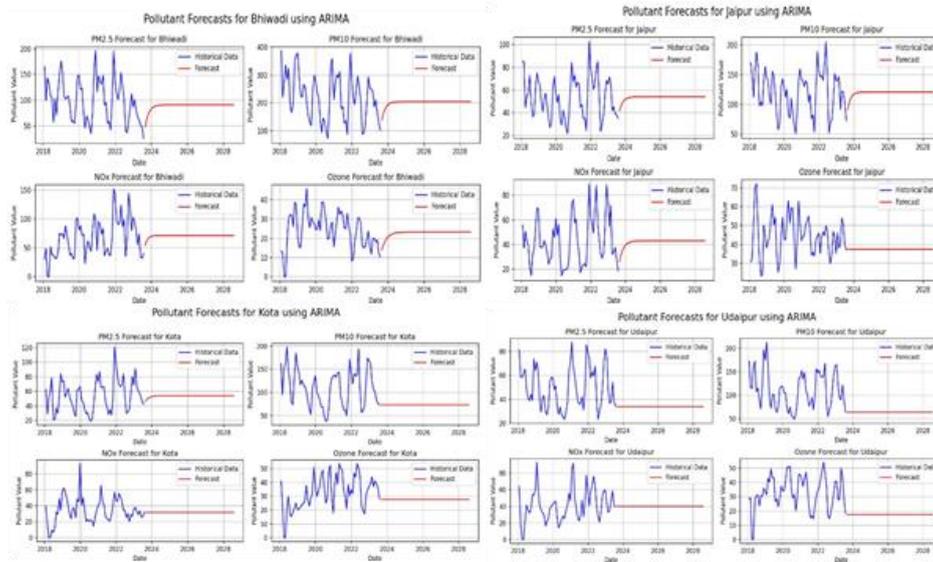

Fig 6: Forecasting of Pollutants in Chosen cities using ARIMA Algorithm

B. Linear Regression:

A machine-learning approach that uses supervised learning to perform a regression job is known as Linear Regression. Linear regression can be used as a basic machine learning algorithm for air quality forecasting, particularly when historical data on air quality and relevant features are available [8]. Using this method, a linear equation was fitted to the observed data to ascertain the relationship between a dependent variable and one or more independent variables. The equation for simple linear regression with time series data is

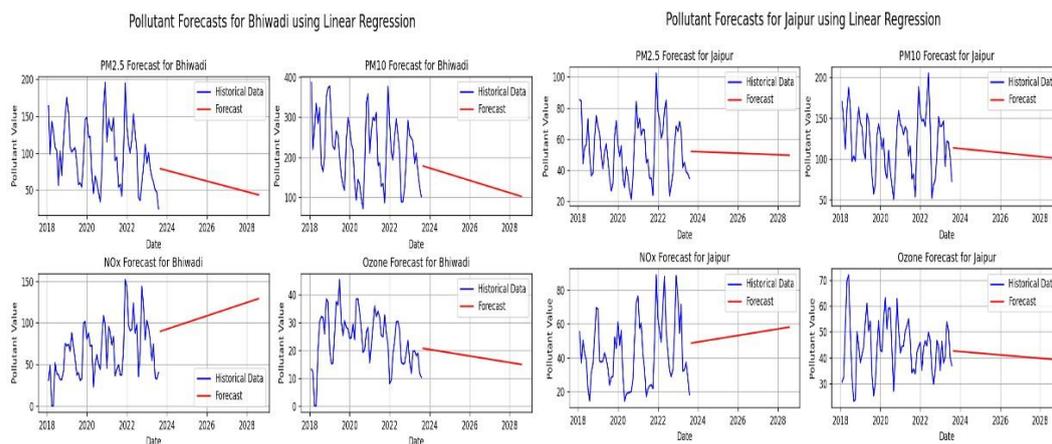$$Y_t = \beta_0 + \beta_0 . X_t + \epsilon_t \quad \text{-----------(1)}$$

Where:

Y_t is the dependent variable at time t, X_t is the independent variable at time t, β_0 is the intercept (y-intercept),

β_1 is the slope of the regression line,

ϵ_t is the error term at time t, representing the difference between the observed and predicted values.

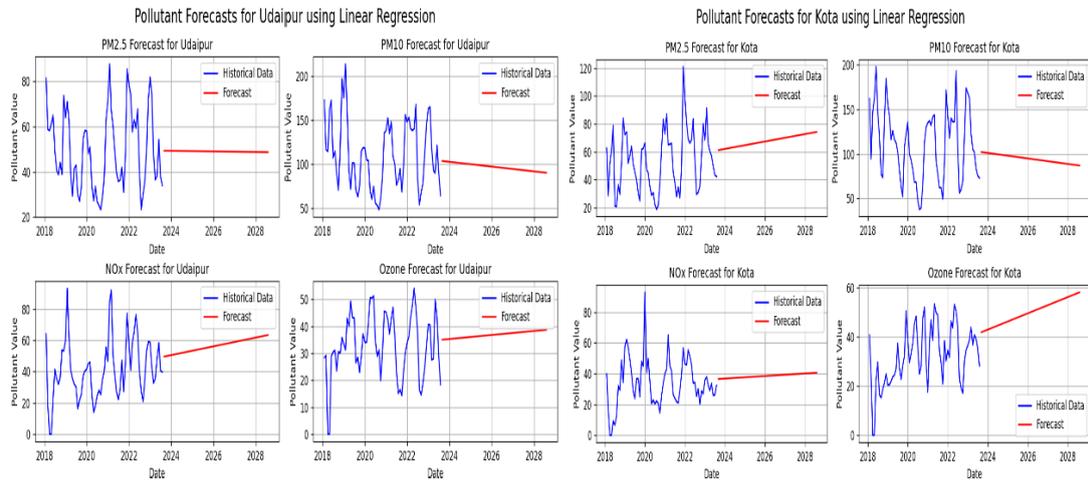The forecasting graph for the next five years in the selected cities is shown below:

Fig 7: Forecasting of Pollutants in Chosen cities using Linear Regression Algorithm

C. Support Vector Machine:

The supervised learning models known as support vector machines (SVMs), or support vector networks, evaluate data, spot trends, and carry out regression and classification tasks. Support vector machines (SVMs) are supervised learning models for classification and regression analysis that make use of matching learning methods. A non-probabilistic binary linear classifier builds a model by using SVM training to classify new data measurements into one of two groups [27]. Forecasting using SVM for the next five years is as follows:
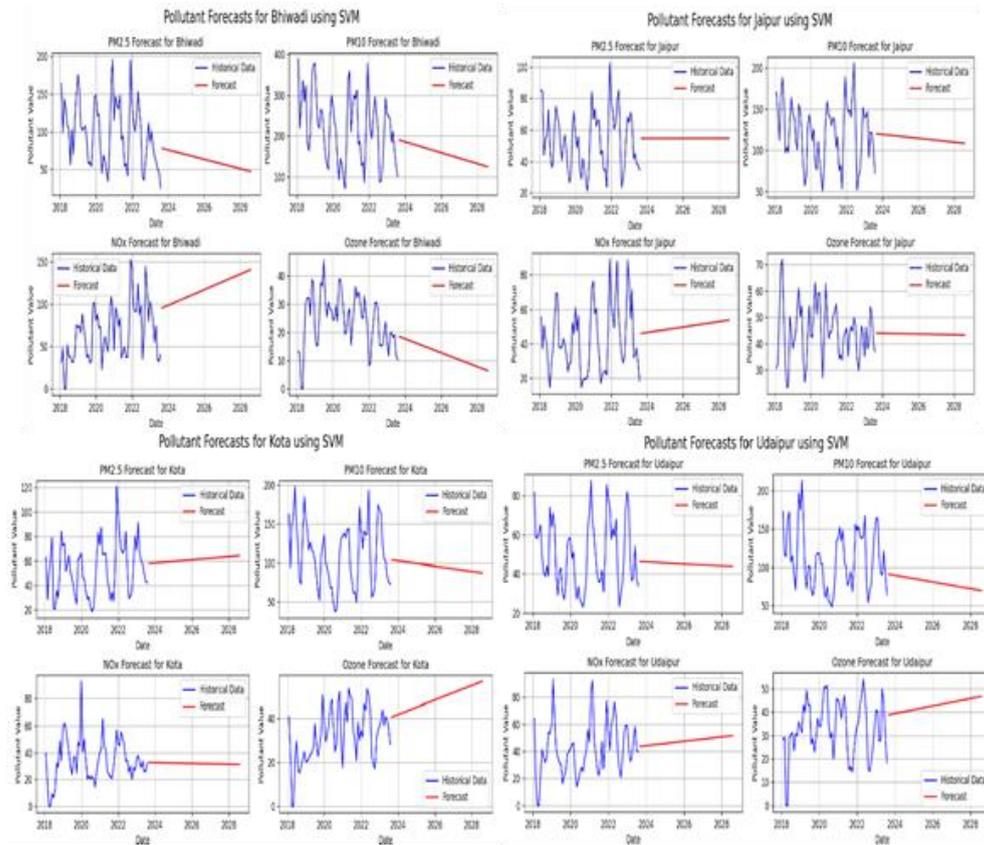

Fig 8: Forecasting of Pollutants in Chosen cities using SVM Algorithm

Exponential Smoothing:

A time-series forecasting technique called exponential smoothing gives historical observations exponentially diminishing weights. It is especially helpful for predicting when data exhibits seasonality or a trend. An appropriate technique for air quality data that may exhibit temporal patterns and trends is exponential smoothing. The forecasting results for the five years obtained using Exponential Smoothing are given below:
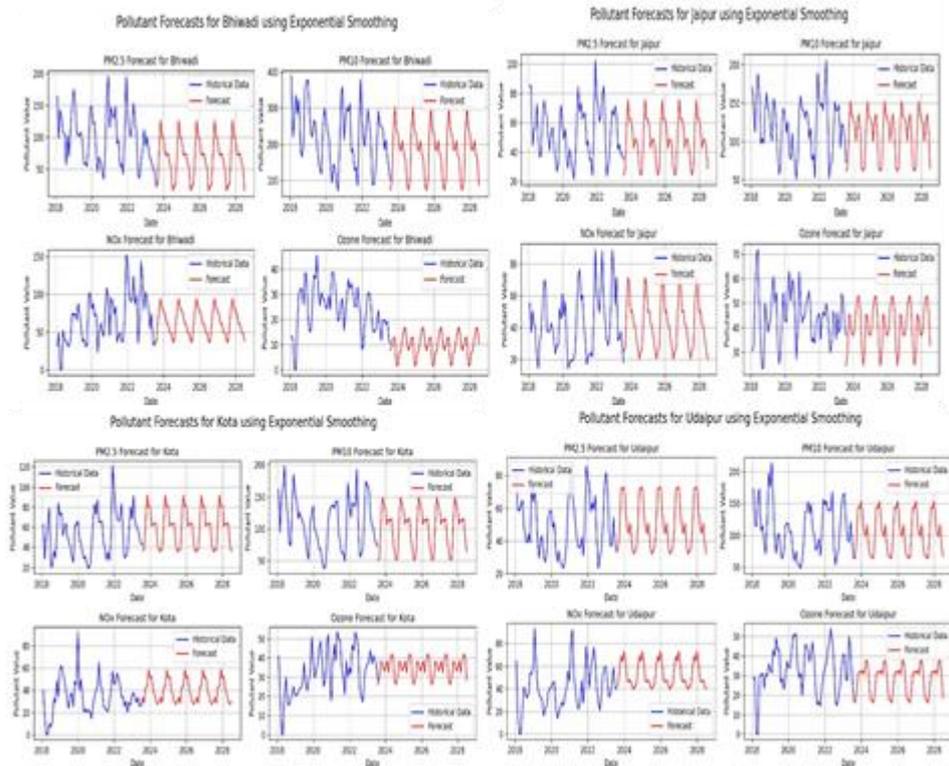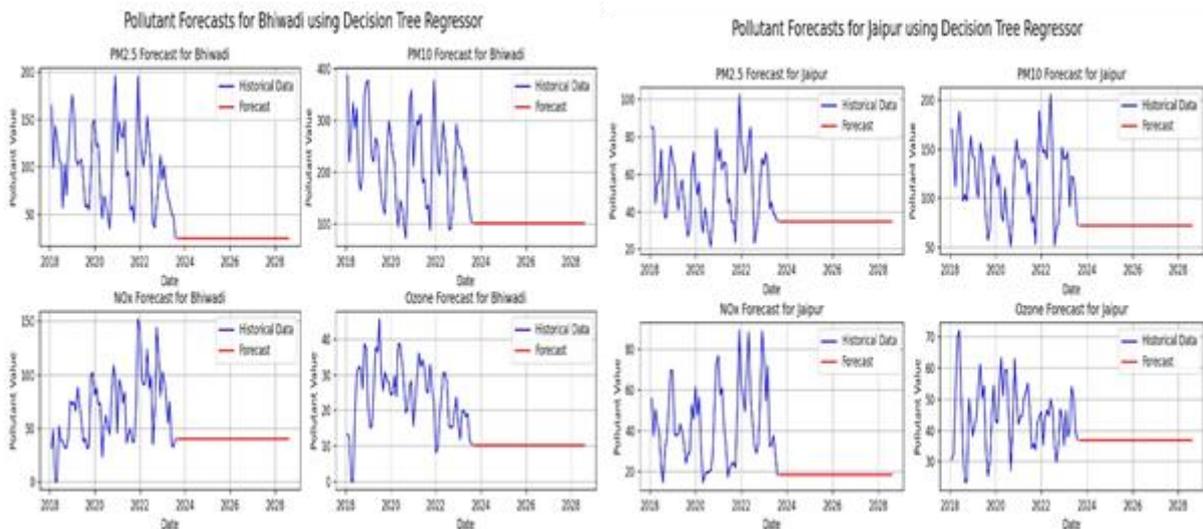
Fig 9: Forecasting of Pollutants in Chosen cities using Exponential Smoothing

D. Decision Tree:

The Decision Tree method is a machine learning tool that recursively divides the data into subsets according to the features, making decisions at each node of the tree. It is primarily used for classification problems and time-series data. A decision Tree offers a nonparametric technique for partitioning datasets [26].

Decision Tree was selected as one of the algorithms for this comparative study as it can transform enormous, complex datasets into graphical displays that are both easy to grasp and packed with information. The forecasting results obtained after applying the decision tree are as follows.
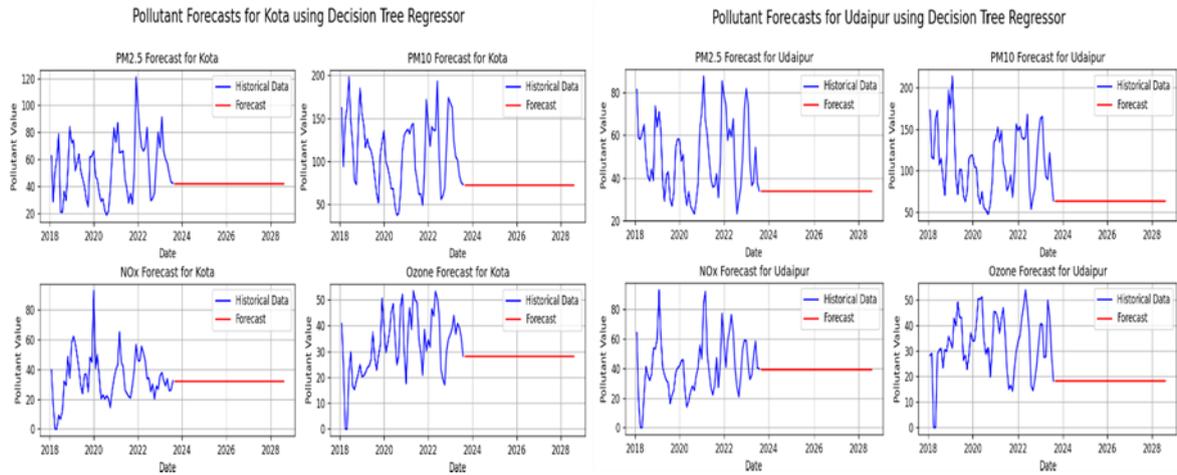
Fig 10: Forecasting of Pollutants in Chosen cities using Decision Tree Regressor

E. XG Boost:

Using a set of techniques, several weak learners can be transformed into a strong prediction model, a process known as "boosting." One potent boosting method that has shown success in a variety of prediction applications, including air quality forecasting, is Extreme Gradient Boosting (XG Boost). An ensemble learning technique called XG Boost constructs a number of weak learners, usually decision trees, and then aggregates their predictions to create a reliable and accurate model [18]. The forecasting results obtained using the XG Boost are as follows:
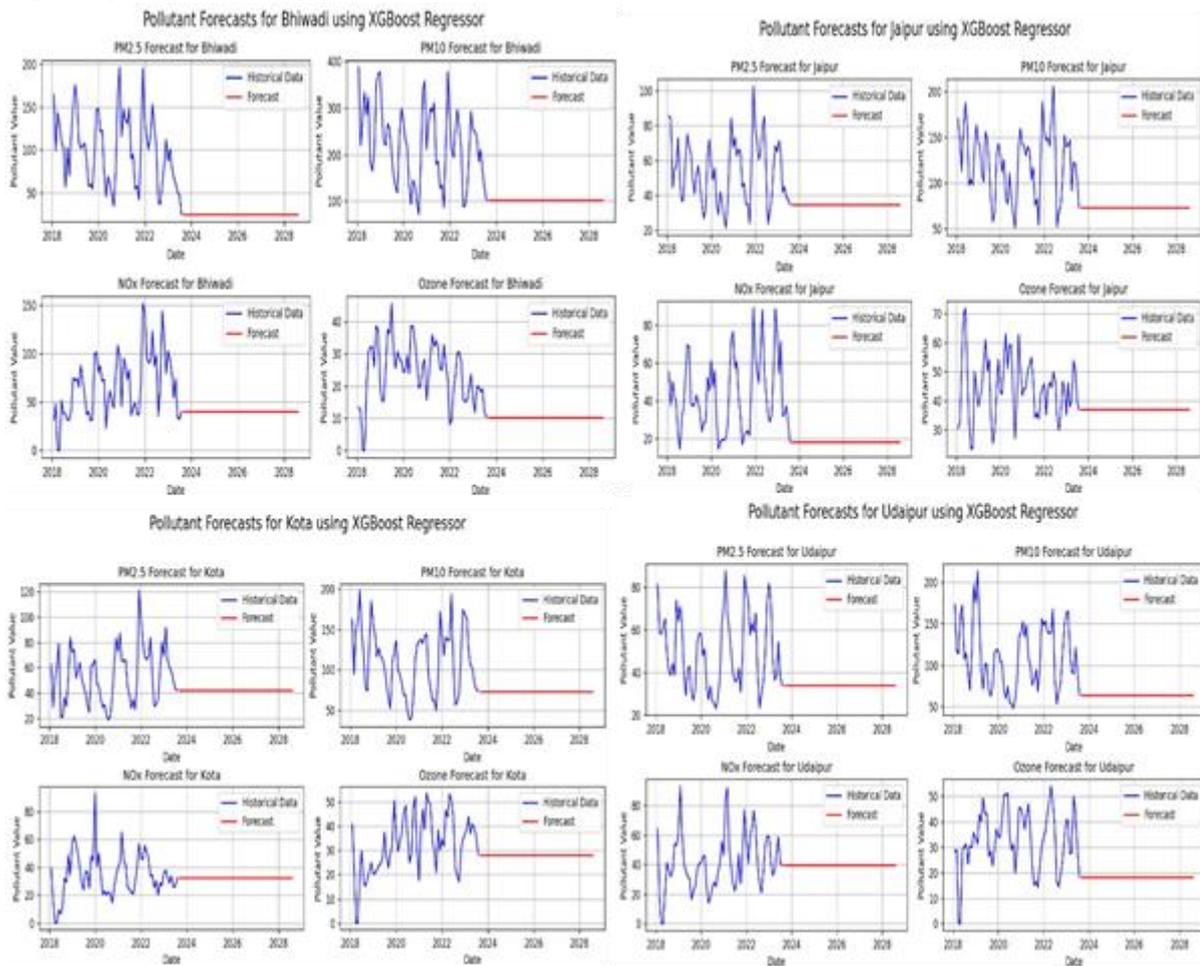

Fig 11: Forecasting of Pollutants in Chosen cities using XG Boost Technique

F. Random Forest:

A collection of decision trees called random forests was used for classification and regression. The majority vote was ascertained through the classification. Regression analysis was used to calculate the mean values. In addition to being more reliable and accurate, this approach can handle a variety of data types such as

continuous, category, and binary data. A group of decision trees is a Random Forest is. Random forest is a meta-estimator that combines and enhances many decision trees by integrating the outcomes of multiple forecasts. A percentage of the total, referred to as the hyper parameter, limits the number of functions that can be split at each node [6]. Random Forest is considered for the comparative study because it has features such as Handling Non-Linearity, Robust to Overfitting, Feature Importance, and Robustness to outliers. The forecasting results are as follows:
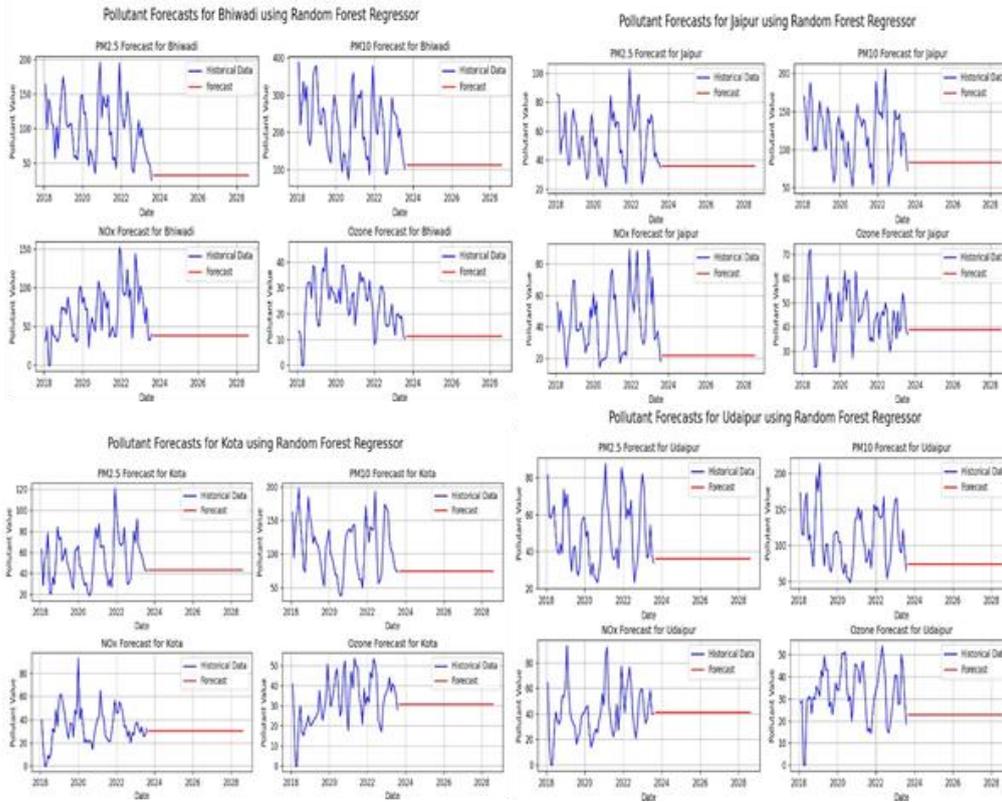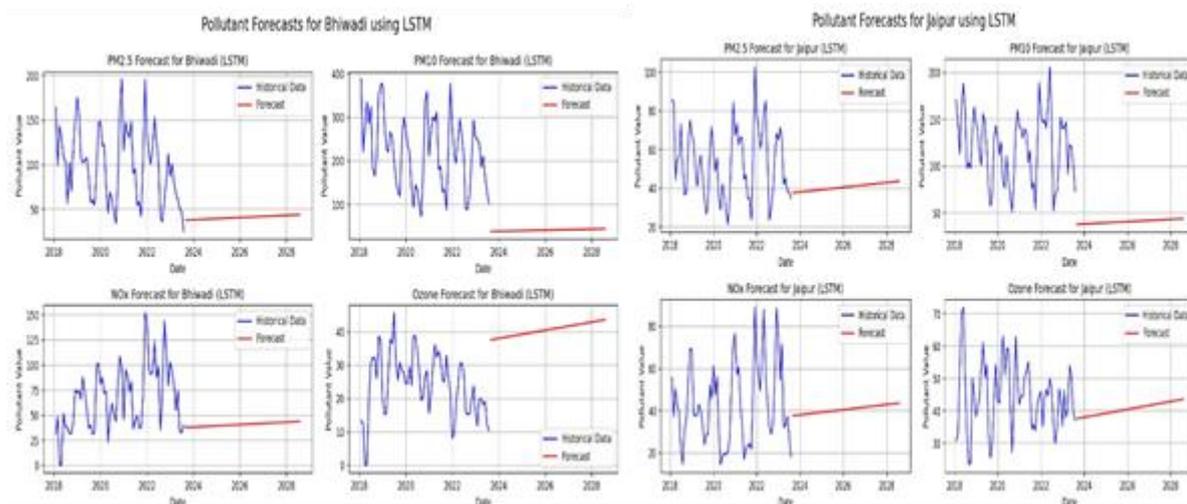

Fig 12: Forecasting of Pollutants in Chosen cities using Random Forest Regressor

Long Short-Term Memory:

Special recurrent neural networks, known as long- and short-term memory (LSTM) neural networks, are able to learn dependent information over an extended period of time and successfully prevent the phenomena of a fading gradient. The machine learning architecture allows the model to "learn" across an extensive number of time steps. By establishing a connection between the memory cell and the neural nodes of the hidden layer of the cyclic neural network, it is also capable of recording historical information. This can be achieved by incorporating three gate structures: input, forget, and output [25]. The forecasting results obtained using LSTM are as follows:
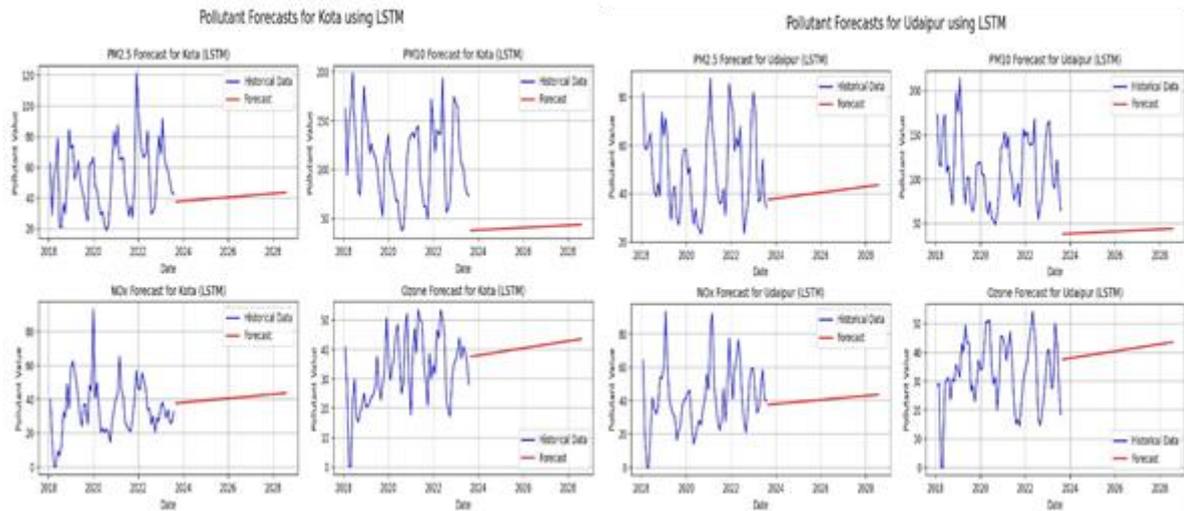
Fig 13: Forecasting of Pollutants in Chosen cities using LSTM

5. Comparison:

To determine the set of rules that performed exceptionally well, the test results were examined and compared. We trained eight models: ARIMA, LR, SVM, Exponential Smoothing, Decision Tree Regressor, XG Boost, Random Forest and LSTM. We carried out the training and testing of the model in Python using the Jupyter Notebook and forecasted the concentration of pollutants for the next five years, as depicted in Figures 6 to 13. After the forecasting, we calculated the error rates for each algorithm.

Air-quality forecasting is a regression problem. The performance of the algorithms used in regression problems was evaluated using regression metrics, also known as error rates. The error rates used were (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). For air quality forecasting in the selected areas of Rajasthan, these metrics help us assess how well the model predicts pollutant concentrations on air quality indices. The different error rates obtained for each algorithm are shown in Figure 14.

Fig 14: Average Error Metrics of Each Algorithm
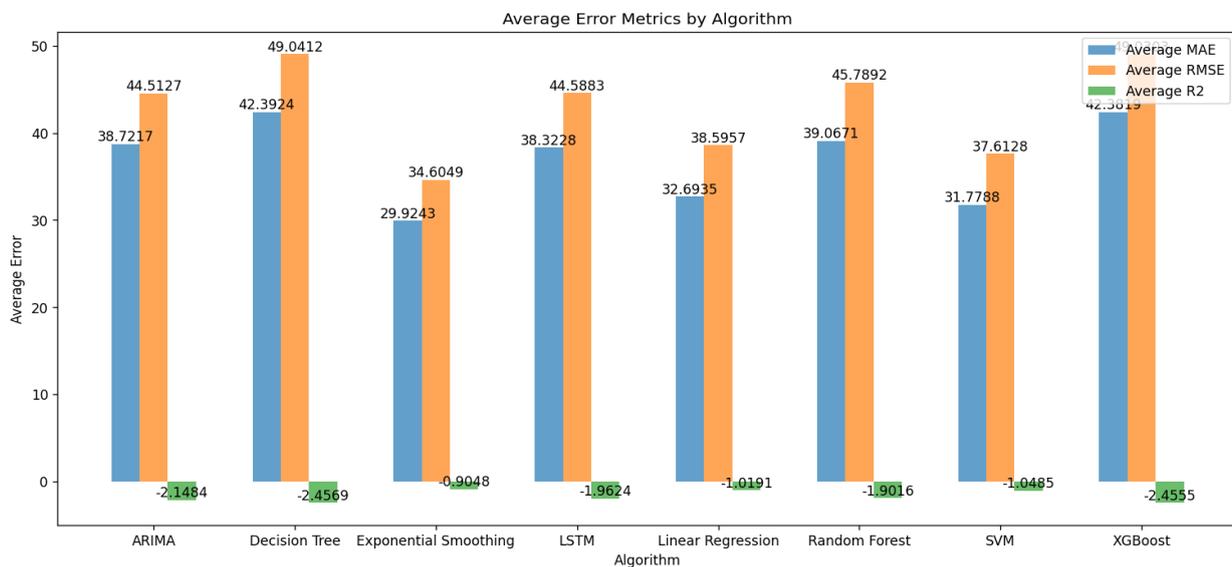
After calculating the error rates, the accuracy of each algorithm was calculated as follows: How accurately each algorithm predicted the future concentration of each pollutant. The results of how accurately each algorithm forecasted the air pollutant concentration in the 4 locations of Rajasthan, namely, Bhiwadi, Jaipur, Kota, and Udaipur, are shown below in the figure.
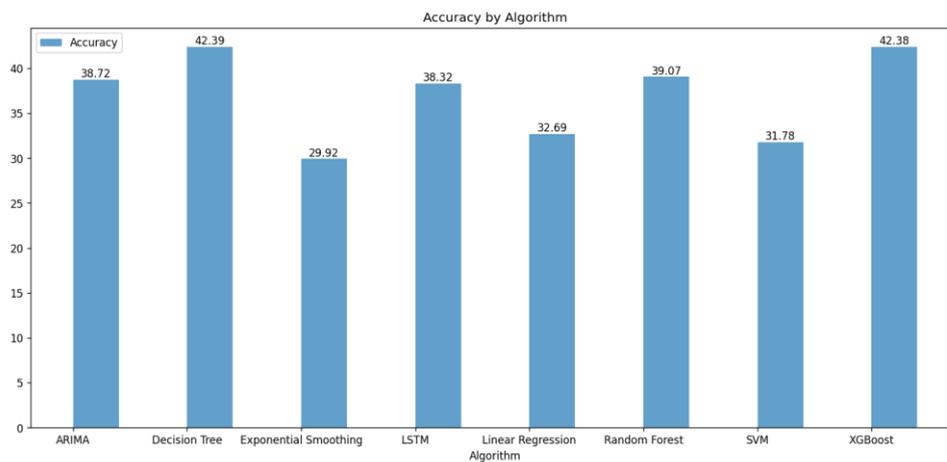
Fig 15: Accuracy Percentage of each Algorithm

which would turn out to give more accurate and The graph above summarizes the results of various algorithms applied to air quality data collected from different monitoring stations. It can be clearly predicted from the graph that the decision tree proved to be the most promising algorithm along with XG Boost with 42.39% and 42.38% accuracy, respectively, for predicting air quality in the highly polluted urban centres of Rajasthan. The next set of algorithms that gave satisfactory results were Random Forest with 39.07% of accuracy, ARIMA with 38.72% accuracy, and LSTM with 38.32 % accuracy. The remaining algorithms did not provide satisfactory results; hence, they could be discarded.

Conclusion:

The dynamic environment, unpredictability, and variability of contaminants in chosen locations and time make air quality prediction a difficult endeavour. This research paper delved into the critical domain of air quality forecasting within the four most polluted cities of Rajasthan: Bhiwadi, Jaipur, Kota, and Udaipur. Leveraging a diverse set of algorithms, including ARIMA, Random Forest, LSTM, Exponential Smoothing, XG Boost, Decision Tree, and SVM, the study aimed to predict concentration of the most prevalent pollutants in these cities which are PM2.5, PM10, NOx, and Ozone. For assessing the accuracy of these algorithms different error metrics were utilized and the results were calculated as depicted in Figure 14. After the calculating the accuracy, the outcomes revealed that Decision Tree and XG Boost outperformed other algorithms, underscoring their efficacy in accurately predicting air quality in the specified urban locales. This research contributes significantly to the understanding and application of different statistical, Machine learning and Neural Network techniques for precise air quality forecasting, offering valuable insights for environmental monitoring and strategic interventions in the highly polluted regions of Rajasthan. In future for making a

more efficient model for air quality forecasting like a Hybrid or an ensemble model these algorithms can be considered

promising results which would definitely help the Government and Environment Organization for policy making and taking necessary actions to control pollution in Rajasthan.

References:

[1] Bhatti, U. A., Yan, Y., Zhou, M., Ali, S., Hussain, A., Qingsong, H., ... & Yuan, L. (2021). Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach. Ieee Access, 9, 41019- 41031.

[2] Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California. Complexity, 2020.

[3] Chen, X., Yin, L., Fan, Y., Song, L., Ji, T., Liu, Y.,... & Zheng, W. (2020). Temporal evolution characteristics of PM2. 5 concentrations based on continuous wavelet transform. Science of The Total Environment, 699, 134244.

[4] Du, S., Li, T., Yang, Y., & Horng, S. J. (2019).

[5] Deep air quality forecasting using hybrid deep learning framework. IEEE Transactions on Knowledge and Data Engineering, 33(6), 2412- 2424.

[6] Freeman, B. S., Taylor, G., Gharabaghi, B., & Thé,

[7] J. (2018). Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association, 68(8), 866-886.

[8] Gochhait, S. A. I. K. A. T., Rimal, Y. A. G. Y. A. N.

[9] T. H., & Pageni, S. A. K. U. N. T. A. L. A. (2021).

The comparison of forward and backward neural network model–A study on the prediction of student grade. WSEAS Transactions on Systems and Control, 16, 422-429

[10] Gopu, P., Panda, R. R., & Nagwani, N. K. (2021). Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India. In Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1 (pp. 47-56). Springer Singapore.

[11] Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. Journal of Environmental and Public Health, 2023

[12] https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard/caaqm-landing/data

[13] Kumar, K., & Pande, B. P. (2023). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology, 20(5), 5333-5348

[14] Lang, P. E., Carslaw, D. C., & Moller, S. J. (2019). A trend analysis approach for air quality network data. Atmospheric Environment: X, 2, 100030.

[15] Liu, H., Yan, G., Duan, Z., & Chen, C. (2021). Intelligent modeling strategies for forecasting air quality time series: A review. Applied Soft Computing, 102, 106957.

[16] Ma, J., Ding, Y., Gan, V. J., Lin, C., & Wan, Z. (2019). Spatiotemporal prediction of PM2. 5 concentrations at different time granularities using IDW-BLSTM. Ieee Access, 7, 107897-107907

[17] Maaloul, K. A. M. E. L., & Brahim, L. E. J. D. E.

[18] L. (2022). Comparative Analysis of Machine Learning for Predicting Air Quality in Smart Cities. WSEAS Transaction on Computers

[19] Mahendra, H. N., Mallikarjunaswamy, S., Kumar,

[20] D. M., Kumari, S., Kashyap, S., Fulwani, S., & Chatterjee, A. (2023). Assessment and Prediction of Air Quality Level Using ARIMA Model: A Case Study of Surat City, Gujarat State, India. Nature Environment & Pollution Technology, 22(1).

[21] Mao, W., Wang, W., Jiao, L., Zhao, S., & Liu, A. (2021). Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. Sustainable Cities and Society, 65, 102567.

[22] Pan, Q., Harrou, F., & Sun, Y. (2023). A comparison of machine learning methods for ozone pollution prediction. Journal of Big Data, 10(1), 63.

[23] Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2(7), 1308.

[24] Saraswat, P., & Singh, S. STUDY OF PARTICULATE MATTER (PM10 AND SPM) DISTRIBUTION IN AMBIENT AIR OF JODHPUR, RAJASTHAN.

[25] Shrivastav, L. K., & Jha, S. K. (2021). A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. Applied Intelligence, 51, 2727-2739.

[26] Talamanova, I., & Pllana, S. (2022, December). Data-driven Real-time Short-term Prediction of Air Quality: Comparison of ES, ARIMA, and LSTM. In International Conference on Intelligent Systems Design and Applications (pp. 322-331). Cham: Springer Nature Switzerland

[27] Verma, A., & Bhatia, D. L. (2023). Analysis of Meteorological Factors Affecting Air Quality in Jaipur Using Time Series Analysis. Available at SSRN 4663529.

[28] Verma, A., & Bhatia, L. Time Series Analysis Using Arima Model for Air Pollution Prediction in Cities of Rajasthan

[29] Xu, S., Li, W., Zhu, Y., & Xu, A. (2022). A novel hybrid model for six main pollutant concentrations forecasting based on improved LSTM neural networks. Scientific Reports, 12(1), 14434

[30] Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., ... & Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. The Innovation, 2(4).

[31] Yang, Y., Farid, S. S., & Thornhill, N. F. (2013). 23 European Symposium on Computer Aided Process Engineering: Prediction of biopharmaceutical facility fit issues using decision tree analysis (Vol. 32). Elsevier Inc. Chapters.

[32] Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308-324.